Ames ...
72-52 ...

252401
12 P.

Bayesian Classification

Peter Cheeseman,* Matthew Self,† John Stutz,‡
James Kelly,† Will Taylor† and Don Freeman†

NASA Ames Research Center
Mail Stop 244-17
Moffett Field, CA 94035

Draft Revised 1:10 AM, March 8, 1988

~3700 Words

Machine Learning: Unsupervised Classification

Abstract

This paper describes a Bayesian technique for unsupervised classification of data and its computer implementation, AutoClass. Given real valued or discrete data, AutoClass automatically determines the most probable number of classes present in the data, the most probable descriptions of those classes, and each object's probability of membership in each class. The program performs as well as or better than existing automatic classification systems when run on the same data, and contains no *ad hoc* similarity measures or stopping criteria. Researchers have also applied AutoClass to several large databases where it has discovered classes corresponding to new phenomena which were previously unsuspected.

(NASA-TM-101274) BAYESIAN CLASSIFICATION (NASA) 12 p N90-70554

Unclas
00/82 0252401

*RIACS. This work partially supported by NASA grant NCC2-428.
†Sterling Software (Don Freeman is now at the University of Pittsburgh)
‡NASA Ames Research Center

1 Introduction

AutoClass is an automatic classification system which calculates probabilistic descriptions of classes in data, rather than merely enumerating the objects in each class. The system implements the Bayesian solution to the problem of separating a finite mixture distribution and provides a sound procedure for determining the number of classes present in such a mixture. Rather than making assumptions about what classes the user desires, The AutoClass approach makes assumptions about the nature of the actual classes and then uses Bayes's theorem to derive the optimal separation criterion. No additional principles are required.

The resulting classification system has several important advantages over most previous work:

- AutoClass automatically determines the most probable number of classes given the data and class model. The classes found represent actual structure in the data. Given random data, AutoClass generates a single class.

- Classification is probabilistic. Class descriptions are given in terms of probability distributions, and Bayes's theorem is all that is required to perform classification. No ad hoc similarity measure, stopping rule, or clustering quality criterion is needed. Decision theory is directly applicable to the probability distributions calculated by AutoClass.

- Class assignments are not absolute. No datum is completely included in or excluded from any class. The resulting "fuzzy" classes capture the common sense notion of class membership better than a categorical classification.

- Real valued and discrete attributes may be freely mixed. Missing attribute values and "tree valued" attributes can be easily incorporated into the AutoClass model also.

- Classifications are invariant to changes of the scale or origin of the data.

2 Theory

When classifying a database, AutoClass does not attempt to partition the data into classes, but rather computes probabilistic descriptions of classes which account for the observed data. In order to find classes in a set of data, we make explicit assumptions about the nature of the classes. These assumptions take the form of parameterized probabilistic models of the classes, where the parameters are unknown. The task of classification then becomes the problem of estimating these classification parameters from a given database. The class distributions are defined over the attribute space of the objects and give the probability distribution of the attributes of an object known to belong to a given class. Classification has long been studied in these terms as the theory of finite mixtures. Everitt and Hand [6] provide an excellent review containing over 200 references.

AutoClass is an implementation of the Bayesian solution to the mixture separation problem. We begin with an uninformative prior probability distribution over the classification parameters (which expreses our *a priori* ignorance of the parameters) and we then update this distribution using the information in the database to calculate the posterior probability distribution of the parameters. This posterior distribution allows us to determine both the most probable classification parameters for any number of classes, and the most probable number of classes present in the database. From this information it is also possible to calculate the probability that each object is a member of each class. Note that it is possible to accurately determine the parameters of strongly overlapping classes despite the fact that very few of the objects can be assigned to any class with high probability.

In addition to providing the database, the user selects an appropriate class model (by defining the class distributions). AutoClass then calculates the optimal values of the parameters for various numbers of classes and the probability that each number of classes is actually present in the data. As final output, AutoClass provides the most probable number of classes, the most probable values of the classification parameters for that number of classes, and also the probability of membership of each object in each class.

AutoClass uses a Bayesian variant of Dempster and Laird's EM Algorithm [3] to search for the maximum of the posterior distribution of the classification parameters and forms an approximation to the distribution about this maximum. AutoClass also includes heuristic tecniques for avoiding local maxima in the search. Although these computational issues are quite difficult to solve in practice, they are only algorithmic issues and do not require any additional theory. Greater details of the Bayesain theory of finite mixtures is given in the Appendix. The AutoClass algorithm is descirbed thoroughly by Cheeseman *et al.* [1]

The class descriptions produced by AutoClass can be used for prediction of future objects. For example, if AutoClass is given a database of symptoms and diagnosed diseases, AutoClass can find classes which can be used to predict the disease of a new object given its symptoms. This prediction is optimal given the assumptions about the causal mechanisms expressed in the class distributions.

AutoClass can also be used to learn from examples. Objects may be presented to AutoClass pre-classified by a teacher. Thus tutored learning can be combined with untutored learning in the same system and using the same theory.

3 Assumptions

The major assumption of AutoClass (and any mixture method) is that a family of class distributions can be found which account for the observed data. AutoClass treats the class distributions modularly so the user is free to develop new class distributions—the user is not constrained to use the class distributions supplied with AutoClass.

The current AutoClass program (AutoClass II) assumes that all attributes are independent within each class. Discrete attributes can take on arbitrary multinomial distributions, and real valued attributes are assumed to be distributed normally. The model does permit any attribute values to be missing from the data. Despite these restrictive assumptions, AutoClass II is able to discern structure in many actual data bases, as discussed in Section 4.

We have nearly completed AutoClass III which includes multivariate normal distributions and exponential distributions for real attributes. We are also developing the theory for automatic selection of class distributions, allowing the system to take advantage of increased model complexity when the data justify the estimation of the additional parameters. Thus, simple theories (with correspondingly few parameters) can give way to more complex theories as the amount of data increases. The theory for such model selection is very similar to the selection of the number of classes.

It is important to point out that we do *not* assume that the classification parameters or the number of classes are "random variables." Rather, we merely assume that they are unknown quantities about which we wish to perform inference. The prior distributions used do not represent a frequency distribution of the parameters, but rather the state of knowledge of the observer (in this case AutoClass) before the data are observed. Thus there can be no "true values of the prior probabilities" as Duda and Hart suggest [5], since prior probabilities are a function of the observer, not of the world. Although Cox gave the first full explanation of this issue in 1946 [2], it remains a source of confusion today.[1]

Bayesian methods have often been discredited due to their use of prior distributions, and the belief that this implies their results are personalistic and therefore somewhat arbitrary. The default prior distribution used in AutoClass, however, is uninformative and completely impersonal.[2] It is invariant to any change of scale or origin, so in no way does it express any *a priori* opinions or biases. Rather, it expresses complete *a priori* ignorance of the parameters (as defined by specific invariance criteria). On the other hand, the ability to incorporate prior knowledge can be of great use when such information is available. Many non-Bayesian approaches have difficulty incorporating such information directly. AutoClass provides the user with the option of incorporating prior information into the classification or using the uninformative prior distribution.

4 Applications

Autoclass has classified data supplied by researchers active in various domains and has yielded some new and intriguing results:

- **Iris Database**

[1]See Jaynes [9] for a recent discussion of the nature of Bayesian inference and its relationship to other methods of statistical inference.

[2]See Jaynes [11] for a lucid description of uninformative priors.

Fisher's data on three species of iris [8] are a classic test for classification systems. AutoClass discovers the three classes present in the data with very high confidence, despite the fact that not all of the cases can be assigned to their classes with certainty. Wolfe's NORMIX and NORMAP [15] both incorrectly found four classes, and Dubes's MH index [4] offers only weak evidence for three clusters.

- **Soybean Disease Database**

  AutoClass found the four known classes in Stepp's soybean disease data, providing a comparison with Michalski's CLUSTER/2 system [13]. AutoClass's class assignments exactly matched Michalski's—each object belonged overwhelmingly to one class, indicating exceptionally well separated classes for so small a database (47 cases, 35 attributes).

- **Horse Colic Database**

  AutoClass analyzed the results of 50 veterinary tests on 259 horses and extracted classes which provided reliable disease diagnoses, despite the fact that almost 40% of the data were missing.

- **Infrared Astronomy Database**

  The Infrared Astronomical Satellite tabulation of stellar spectra is not only the largest database Autoclass has assayed (5,425 cases, 94 attributes) but the least thoroughly understood by domain experts. AutoClass's results differed significantly from NASA's previous analysis. Preliminary evaluations of the new classes by infrared astronomers indicate that the hitherto unknown classes have important physical meaning. The AutoClass infrared source classification is the basis of a new star catalog to appear shortly.

We are actively collecting and analyzing other databases which seem appropriate for classification, including an AIDS database and a second infrared spectral database.

5 Comparison with Other Methods

Several different communities are interested in automatic classification, and we compare AutoClass to some existing methods:

- **Maximum Likelihood Mixture Separation**

  AutoClass is most similar to the maximum likelihood methods used to separate finite mixtures as described in the statistical pattern recognition literature. The mathematical statement of the problem is identical to that discussed by Duda and Hart [5], and by Everitt and Hand [6], The primary difference lies in AutoClass's Bayesian formulation, which provides a more effective method for determining the number of classes than existing methods based on hypothesis testing. A more detailed comparison of AutoClass to maximum likelihood methods is given by Cheeseman *et al.* [1]

- **Cluster Analysis**

  Cluster analysis and AutoClass's finite mixture separation differ fundamentally in their goals. Cluster analysis seeks classes which are groupings of the data points, definitively assigning points to classes; AutoClass seeks descriptions of classes that are present in the data, and never assigns points to classes with certainty.

  The other major difference lies in the assumptions made about the form of the classes. To attempt the problem of classification, some assumptions must be made about the nature of the classes sought. The AutoClass method makes these assumptions directly by specifying class distributions and then derives the optimal class separation criterion using Bayes's theorem. Cluster analysis techniques make their assumptions indirectly by specifying a criterion for evaluating clustering hypotheses, such as maximizing intra-class similarity.

- **Conceptual Clustering**

  Both AutoClass and conceptual clustering methods seek descriptions of the clusters rather than a simple partitioning of the objects. The main difference between the methods is the choice of concept language: AutoClass uses a probabilistic description of the classes, while

Michalski and Stepp [14] use a logical description language. The logic-based approach is particularly well suited to logically "clean" applications, whereas AutoClass is effective when the data are noisy or the classes overlap substantially.

Conceptual clustering techniques specify their class assumptions with a "clustering quality criterion" such as Fisher's category utility [7]. As with cluster analysis, these are assumptions about what clusterings are desired rather than about the nature of the actual clusters. This may reflect a difference in goals since Langley's CLASSIT [12] and Michalski's CLUSTER/2 [13] programs seek explicitly to emulate human classification, which is a more difficult problem than AutoClass addresses.

6 Conclusion

We have developed a practical and theoretically sound method for determining the number of classes present in a mixture, based solely on Bayes's theorem. Its rigorous mathematical foundation permits the assumptions involved to be stated clearly and analyzed carefully. The AutoClass method performs better at determining the number of classes than existing mixture separation methods and also compares favorably with cluster analysis and conceptual clustering methods.

Appendix

This appendix presents the Bayesian theory of finite mixtures. This theory is the mathematical basis of the AutoClass algorithm.

In the theory of finite mixtures, each datum is assumed to be drawn from one of m mutually exclusive and exhaustive classes. Each class is described by a *class distribution*, $p(x_i \mid x_i \in C_j, \vec{\theta}_j)$, which gives the probability distribution of the attributes of a datum if it were known to belong to the class C_j. These class distributions are assumed to be parameterized by a *class parameter vector*, $\vec{\theta}_j$, which for a normal distribution would consist of the class mean, μ_j, and variance, σ_j^2. The probability of an object being drawn from class j is called the ***class probability*** or mixing proportion, π_j. Thus, the probability distribution of a datum drawn from a mixture distribution is

$$p(x_i \mid \vec{\theta}, \vec{\pi}, m) = \sum_{j=1}^{m} \pi_j \, p(x_i \mid x_i \in C_j, \vec{\theta}_j). \tag{1}$$

We assume that the data are drawn from an exchangeable (static) process—that is, the data are unordered and are assumed to be independent given the model. Thus, the *joint* probability distribution of a set of n data drawn from a finite mixture is

$$p(\vec{x} \mid \vec{\theta}, \vec{\pi}, m) = \prod_{i=1}^{n} p(x_i \mid \vec{\theta}, \vec{\pi}, m). \tag{2}$$

For a given value of the class parameters, we can calculate the probability that an object belongs to each class using Bayes's theorem,

$$p(x_i \in C_j \mid x_i, \vec{\theta}, \vec{\pi}, m) = \frac{\pi_j \, p(x_i \mid x_i \in C_j, \vec{\theta}_j)}{p(x_i \mid \vec{\theta}, \vec{\pi}, m)}. \tag{3}$$

Thus, the classes are "fuzzy" in the sense that even with perfect knowledge of an object's attributes, it will only be possible to determine the probability that it is a member of a given class.

We break the problem of identifying a finite mixture into two parts: determining the classification parameters for a given number of classes, and determining the number of classes. Rather than seeking an ***estimator*** of the classification parameters (the class parameter vectors, $\vec{\theta}$, and the class probabilities, $\vec{\pi}$), we seek their full posterior probability distribution. The posterior distribution is proportional to the product of the prior distribution of the parameters, $p(\vec{\theta}, \vec{\pi} \mid m)$, and the likelihood function, $p(\vec{x} \mid \vec{\theta}, \vec{\pi}, m)$:

$$p(\vec{\theta}, \vec{\pi} \mid \vec{x}, m) = \frac{p(\vec{\theta}, \vec{\pi} \mid m) \, p(\vec{x} \mid \vec{\theta}, \vec{\pi}, m)}{p(\vec{x} \mid m)}, \tag{4}$$

where $p(\vec{x} \mid m)$ is simply the normalizing constant of the posterior distribution, and is given by

$$p(\vec{x} \mid m) = \iint p(\vec{\theta}, \vec{\pi} \mid m) \, p(\vec{x} \mid \vec{\theta}, \vec{\pi}, m) \, d\vec{\theta} \, d\vec{\pi}. \tag{5}$$

To solve the second half of the classification problem (determining the number of classes) we calculate the posterior distribution of the number of classes, m. This is proportional to the product of the prior distribution, $p(m)$, and the pseudo-likelihood function, $p(\vec{x} \mid m)$,

$$p(m \mid \vec{x}) = \frac{p(m)\, p(\vec{x} \mid m)}{p(\vec{x})}. \tag{6}$$

The pseudo-likelihood function is just the normalizing constant of the posterior distribution of the classification parameters (Equation 5). Thus, to determine the number of classes, we first determine the posterior distribution of the classification parameters for each possible number of classes. We then marginalize (integrate) out the classification parameters from the estimation of the number of classes—in effect, treating them as "nuisance" parameters.

In general, the marginalization cannot be performed in closed form, so AutoClass searches for the maximum of the posterior distribution of the classification parameters (using a Bayesian variant of Dempster and Laird's EM Algorithm [3]) and forms an approximation to the distribution about this maximum. See Cheeseman *et al.* [1] for full details of the AutoClass algorithm.

References

[1] Peter Cheeseman, Don Freeman, James Kelly, Matthew Self, John Stutz, and Will Taylor. Autoclass: a Bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning*, Forthcoming.

[2] R. T. Cox. Probability, frequency, and reasonable expectation. *American Journal of Physics*, 17:1–13, 1946.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[4] Richard C. Dubes. How many clusters are best? — an experiment. *Pattern Recognition*, 20(6):645–663, 1987.

[5] Richard O. Duda and Peter E. Hart. *Pattern Recognition and Scene Analysis*, chapter 6. Wiley-Interscience, 1973.

[6] B. S. Everitt and D. J. Hand. *Finite Mixture Distributions. Monographs on Applied Probability and Statistics*, Chapman and Hall, London, England, 1981. Extensive Bibliography.

[7] D. H. Fisher. Conceptual clustering, learning from examples, and inference. In *Proceedings of the Fourth International Workshop on Machine Learning*, pages 38–49, Morgan Kaufmann, 1987.

[8] R. A. Fisher. Multiple measurments in taxonomic problems. *Annals of Eugenics*, VII:179–188, 1936.

[9] Edwin T. Jaynes. Bayesian methods: general background. In James H. Justice, editor, *Maximum Entropy and Bayesian Methods in Applied Statistics*, pages 1–25, Cambridge University Press, Cambridge, Massachusetts, 1986.

[10] Edwin T. Jaynes. *Papers on Probability, Statistics and Statistical Physics*. Volume 158 of *Synthese Library*, D. Reidel, Boston, 1983.

[11] Edwin T. Jaynes. Prior probabilities. *IEEE Transactions on Systems and Cybernetics*, SSC-4(3):227–241, September 1968. (Reprinted in [10]).

[12] Pat Langley, John H. Gennari, and Wayne Iba. Hill-climbing theories of learning. In *Proceedings of the Fourth International Workshop on Machine Learning*, pages 312–323, Morgan Kaufmann, 1987.

[13] Ryszard S. Michalski and Robert. E. Stepp. Automated construction of classifications: conceptual clustering versus numerical taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5:396–410, 1983.

[14] Ryszard S. Michalski and Robert E. Stepp. Learning from observation: conceptual clustering. In Ryszard S. Michalski, Jaime G. Carbonell, and Tom M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach*, chapter 11, Tioga Press, Palo Alto, 1983.

[15] John H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioural Research*, 5:329–350, July 1970.